Nuove risorse per la ricerca del lessico del patrimonio culturale corpora multilingue LBC

Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC

Riccardo Billero, María Carlota Nicolás Martínez
Università degli Studi di Firenze

As a result of the wealth of artistic expression that Italy has produced over the centuries, the Italian cultural heritage lexicon has become a crucial object of interest for scholars of varied disciplines. However, while monolingual art dictionaries are currently available, there are no multilingual tools that offer the same level of quality and comprehensiveness. The present work moves in that direction. In particular, the aim of this paper is to describe the work done to date in designing and implementing the LBC database, a resource for constructing a multilingual art dictionary in nine languages (Chinese, French, English, German, Italian, Portuguese, Russian, Spanish and Turkish). This database, made up of nine corresponding corpora, will contain texts whose subject is cultural heritage, ranging from technical texts on art history to books on art appreciation, such as tour guides, and, lastly, travel books highlighting Italian art and culture. Below is a summary of the decisions taken during the work process and the challenges that lie ahead for its future development.

Keywords: corpus linguistics, lexicography, cultural heritage

1. Introduzione

Il lessico italiano dei beni culturali è un patrimonio che va ben oltre i confini delle comunità linguistiche: esso è infatti oggetto di interesse sia da parte di studiosi afferenti a diverse discipline (storia dell'arte, letteratura, linguistica) sia da parte di professionisti (traduttori, operatori turistici, organizzatori di eventi) e studenti in formazione nelle medesime discipline o professioni. Mentre sono disponibili dizionari monolingue della lingua dell'arte, non sono invece disponibili strumenti multilinguistici aventi lo stello livello di qualità. Pertanto, nel 2013, è stata costituita presso il Dipartimento di Lingue, Letterature e Studi Interculturali dell'Università degli Studi di Firenze l'Unità di Ricerca Lessico dei Beni Cul-

CHIMERA. Romance Corpora and Linguistic Studies 4.2 (2017), 203-216. ISSN 2386-2629

turali (LBC), che mediante il progetto omonimo si pone la finalità di colmare questa lacuna.

Obiettivi fondamentali di tale progetto sono:

- realizzare una piattaforma web di riferimento per l'uso e lo studio del lessico attinente ai beni culturali;
- creare una banca dati con nove corpora delle nove lingue coinvolte nel progetto (cinese, francese, inglese, italiano, portoghese, russo, spagnolo, tedesco e turco);
- offrire, insieme alla banca dati multilingue, strumenti per l'interrogazione dei corpora consentendo e promovendo l'attivazione di aree di ricerca in diversi campi, quali studi linguistici, lessicali, socio-culturali;
- fornire l'accesso ai testi contenuti nei corpora presenti all'interno della piattaforma ad ogni utente registrato[1] e attivando così un dialogo con chiunque voglia contribuire in maniera costruttiva ad arricchire tale piattaforma;
- redigere un dizionario multilingue[2] di termini artistici utilizzando la banca dati come risorsa.

Sebbene il presente lavoro sia partito dallo studio del lessico dei beni culturali, i componenti dell'unità lo affrontano con una visione culturale a ampio spettro ed una vasta apertura al mondo esterno. Infatti, nonostante il presente lavoro abbia avuto origine dalla somma degli interessi culturali e scientifici dei componenti l'unità, legati da ambiti di ricerca ben distinti tra loro (in particolare linguisti, studiosi di lessicografia, specialisti di linguistica computazionale, studiosi di storia dell'arte o di letteratura), tali interessi sono tutti confluiti nello studio del lessico dei beni culturali.

2. Quadro generale

L'impostazione scientifica utilizzata dall'unità di ricerca per quanto riguarda la creazione dei corpora condivide il punto di vista della linguistica dei corpora, qui ben delineata:

[1] È prevista anche la registrazione in base alla tipologia di utenti; a titolo di esempio alcune tipologie di utenti ipotizzati per la piattaforma sono: professore universitario, studioso di lessicografia, studioso storico, studioso artistico, guida turistica, traduttore e studente.

[2] Il rapporto fra corpus e dizionario è trattato accuratamente da Ramos (2009).

> Research in corpus linguistics deals with some set of machine-readable texts which is deemed an appropriate basis on which to study a particular research questions. The set of texts or corpus is usually of a size which defies analysis by hand and eye alone within any reasonable timeframe. For this reason, corpora are invariably exploited using software search tools. Concordancers allow users to look at words in context. Other tools allow the production of frequency data, for example a word frequency list, which lists all words appearing in a corpus and specifies how many times each one occurs in that corpus. Concordances and frequency data exemplify respectively the two forms of analysis, namely qualitative and quantitative, that are equally important to corpus linguistics. (McEnery & Hardie 2012).

Sempre gli stessi autori mostrano che è evidente che ogni studio realizzato su un corpus, e pertanto anche con il corpus oggetto del presente lavoro, segua le linee *corpus-driven* descritte da Tognini-Bonelli (2001: 84-85):

> Corpus-driven linguistics rejects the characterisation of corpus linguistics as a method and claims instead that the corpus itself should be the sole source of our hypotheses about language. It is thus claimed that the corpus itself embodies a theory of language.

Nel nostro progetto, al fine di documentare le voci del dizionario, si useranno delle citazioni provenienti da testi di indiscusso prestigio culturale appartenenti al corpus; oltre al modello *corpus-driven* sarà pertanto utilizzato il modello *corpus-based*, anche questo descritto da Tognini-Bonelli (2001: 84-85):

> The distinction between corpus-based and corpus-driven language study was introduced by Tognini-Bonelli (2001). Corpus-based studies typically use corpus data in order to explore a theory or hypothesis, aiming to validate it, refute it or refine it. The definition of corpus linguistics as a method underpins this approach.

L'abbinamento dei due tipi di approccio sarà presente in molte delle ricerche promosse dall'Unità di ricerca; inizialmente, attraverso gli studi *corpus-driven* sarà possibile fornire lemmari e alberi di rapporto tra i nomi comuni che riguardano un campo delimitato di conoscenza (es. chiese, monumenti, ecc…)[3]. Per esempio, per creare lemma di termini, la *query* semplice "di San??" consente di individuare le chiese e le piazze (Chiesa di Santa Croce, Piazza di Santa Croce, ecc…), mentre una *query* CQL di ricerca dei nomi propri può permettere di discernere fra nomi comuni di luoghi o monumenti (come Battistero o Cupola di

[3] Per l'individuazione dei nomi di luoghi è stato preso come riferimento il lavoro di Tran & Maurel (2006).

Brunelleschi o Piazza Brunelleschi). Invece, per quanto riguarda gli alberi di rapporto, un esempio è dato dalle parole legate[4] alla parola chiesa, quali chiostro, cappella, ecc….

Per quanto riguarda invece l'aspetto informatico, si è deciso di utilizzare strumenti *open source*, poiché:

a. il codice della comunità *open source* è di qualità superiore e rimane affidabile nel tempo;
b. la trasparenza dell'*open source* e il controllo che riceve da tutti gli utenti garantisce una maggiore sicurezza;
c. l'*open source* dispone di ampie comunità di persone che consentono un continuo miglioramento prestazionale e una ampia fonte di risposte ad ogni dubbio;
d. le piattaforme *open source* consentono una maggiore interoperabilità, contro la rigidità di un'unica soluzione che non può appagare tutte le necessità.

3. Scopi specifici della banca dati LBC (Bd-LBC)

La banca dati LBC, contenente i nove corpora, che sarà indicata in breve come Bd-LBC, è una delle risorse fondamentali dell'unità di ricerca LBC; essa è stata progettata (a partire dal 2016) per avere un ampio uso specialistico, prefissandosi tre requisiti fondamentali:

a. disporre di materiale linguistico per documentare le voci del dizionario multilingue LBC;
b. disporre di materiale linguistico per effettuare studi linguistici, letterali o culturali;
c. disporre di testi con i quali far conoscere al grande pubblico il patrimonio culturale di Firenze e della Toscana.

Per quanto riguarda gli utenti, si è tenuto conto fin dall'inizio del fatto che la banca dati non sarà utilizzata solamente dai ricercatori dell'Unità di ricerca LBC[5] nel corso della stesura del dizionario, ma sarà consultabile anche da utenti esterni ad essa, interessati ai contenuti ivi presenti, quali traduttori e specialisti

[4] La caratteristica descritta può essere ottenuta con la funzionalità di word sketch del servizio online SketchEngine.
[5] Un esempio di utilizzo si trova nel lavoro di Carpi (2017).

in storia dell'arte o della cultura italiana[6]. A questi ultimi utenti sarà fornita sia la possibilità di accedere alle singole voci di dizionario, sia di consultare i vari corpora esistenti nelle diverse lingue, al fine di effettuare l'estrazione delle concordanze o di analizzare le liste di frequenze delle varie parole. La registrazione degli utenti e la connessa richiesta di fornire feedback dei lavori svolti utilizzando la nostra banca dati, renderanno possibili analisi sul comportamento degli utenti e permetteranno eventualmente di effettuare modifiche alla banca dati, laddove ritenute necessarie.

4. Tratti definitori e macrostruttura

Occorre considerare quattro diversi parametri, al fine di definire i corpora che compongono la Bd-LBC, ovvero genericità, mezzo di trasmissione, limiti cronologici, lingua (Cresti & Panunzi 2013: 53):

- per quanto riguarda la genericità, i corpora oggetto del presente lavoro devono essere considerati come specialistici, poiché appartengono al linguaggio di un determinato settore culturale;
- rispetto al mezzo di trasmissione, si tratta di testi scritti;
- per quanto riguarda la cronologia, sono corpora diacronici, dal Rinascimento italiano ad oggi;
- infine, ognuno dei corpora è espresso in una sola lingua; questo non impedisce tuttavia che in un secondo tempo possano essere creati corpus paralleli o comparabili[7]. In particolare, la Bd-LBC è composta di nove corpora aventi le stesse caratteristiche per ogni lingua[8].

Per quanto riguarda le scelte fatte a proposito del piano di campionamento, inteso come rappresentatività, si è tenuto conto di aspetti qualitativi e quantitativi.

[6] Attualmente è consultabile in rete un glossario di termini artistici attualmente abbastanza breve, ma che dichiara di essere in espansione, denominato *Glossario dei termini artistici* e realizzato da Finestre sull'Arte, rivista online d'arte antica e contemporanea.

[7] Sebbene attualmente l'italiano abbia un ruolo primario e di riferimento (i testi presenti sono in italiano e nelle loro traduzioni esistenti nelle altre lingue), sono presenti anche testi di riconosciuto prestigio culturale originariamente scritti in altre lingue (come ad esempio *Mornings in Florence* di Ruskin) ed inclusi nella banca dati sia nella loro lingua originale sia nelle traduzioni, quando esistenti.

[8] Sebbene la struttura dei corpora è identica in ogni lingua, è ovvio che i contenuti non lo potranno essere; ad esempio in lingue quali il cinese o il turco non potranno essere presenti alcune delle opere tradotte appartenenti alla tradizione occidentale.

Sebbene i parametri dimensionali non siano stati sinora definiti, si ritiene che un milione di *token* sia il numero minimo per l'esecuzione di *query* di prova. Inoltre, si ipotizza che cinque milioni di *token* sia una quantità raggiungibile in almeno cinque delle lingue coinvolte nel progetto e tale da poter fornire informazioni molto utili, come verificato anche dalla sperimentazione finora eseguita confrontando i nostri testi con lemmari di riferimento. Per quanto riguarda invece il bilanciamento dei componenti, sono state delimitate alcune tipologie testuali, di cui non viene effettuato al momento un bilanciamento. I criteri di selezione dei testi sono stati: la rilevanza storico-culturale dell'opera dell'ambito specifico di studio (ad es. testi di Vitruvio o Leonardo); la diffusione internazionale di un'opera relazionata con l'ambito di studio (es. libri di Vasari); il prestigio dato a livello internazionale al patrimonio italiano da parte di un'opera (es. testi di Stendhal o Ruskin); la specificità dell'argomento in rapporto alla storia dell'arte italiana ed in particolare della Toscana (es. Burckhardt).

La rappresentatività delle risorse è stata ben definita fin dall'inizio attraverso dei criteri di campionamento dei testi scelti che danno fondamento alla struttura del corpus (Cresti & Panunzi 2013: 57).

La macrostruttura frutto del *corpus design*, oltre a soddisfare la rappresentatività del lessico dei beni culturali, viene incontro alle necessità di estrazione di informazioni degli utenti. Per ogni lingua sono previste tre diverse tipologie di estrazione di informazioni:

- carattere cronologico, tenendo conto del periodo di redazione dei testi;
- tipologia del testo, tenendo conto da una parte del tipo di testo dal punto di vista della sua diffusione o tipo di pubblico a cui è diretto, dall'altra considerando il genere testuale (letterario); inoltre sono presenti dizionari, considerati utili per gli studi lessicografici specifici;
- autore dei testi, per permettere di accedere a tutti i testi di un determinato autore.

Possiamo descrivere questi componenti strutturali dei corpora come segue:

- cronologia: i testi contenuti vanno dal Rinascimento ai giorni nostri. Sebbene saranno presenti entrambe le datazioni, l'anno di pubblicazione sarà secondario rispetto a quello di redazione. Quest'ultimo, in particolare, sarà il dato di maggiore interesse per l'estrazione di informazioni, poiché è in un certo qual modo rappresentativo delle caratteristiche linguistiche del periodo considerato; infatti i testi sono stati inseriti nella banca dati rimanendo fedeli all'edizione usata, senza produrre alcun tipo di modernizzazione;

- tipologia di testi: i testi hanno come argomento il patrimonio artistico e il suo lessico ed in particolare un'ampia visione di Firenze e della Toscana descritta da diversi punti di vista, sia tecnico oggettivi che personali soggettivi. In particolare, sono state individuate le seguenti categorie e sottocategorie[9]:

    - Divulgativo
        - Blog
        - Guida
        - Ricettario
        - Rivista
    - Tecnico
        - Architettura
        - Arte
        - Edilizia
        - Enogastronomia
        - Storia
    - Letterario
        - Biografico
        - Fiction
        - Saggistica
    - Dizionario
        - Monolingue
        - Bilingue/plurilingue

- autore e titolo: la struttura della banca dati permette di memorizzare il nome dell'autore e il titolo del testo. È opportuno indicare che sono stati considerati come testi sia libri interi sia frammenti di essi che fossero in qualche modo "autoconclusivi", ad esempio un capitolo di un libro, un articolo di una rivista, ecc… Tale scelta è stata effettuata poiché in molti casi l'intero libro non coincideva con gli interessi del progetto;

[9] Le categorie e sottocategorie individuate dipendono dalla diffusione del genere testuale. Esiste una logica di questa categorizzazione: innanzitutto, il tecnico si differenzia dal divulgativo soprattutto in base al pubblico destinatario, ovvero se esso è un ricettore specializzato o meno. Vi saranno ad esempio guide di arte di fronte ad opere tecniche d'arte, ecc… ed è quindi possibile evidenziare l'esistenza di un parallelismo che si differenzia in base al pubblico.
Inoltre, una terza categoria è data dai dizionari, che sono da considerarsi come testi tecnici nel campo lessico linguistico di nostro interesse.
Infine il genere testuale letterario si rivela utile perché fornisce una visione di Firenze, espressione di come il suo patrimonio viene visto dal punto di vista esterno.

- delimitazione geografica: i testi contenuti hanno come argomento il patrimonio artistico della Toscana; l'ambito geografico cambierà nel futuro ampliandosi a tutta Italia e ad altre culture[10].

5. Flusso di lavoro per la creazione dei corpora

Il flusso di lavoro per l'inserimento di testi all'interno dei corpora è stato condotto da varie squadre linguistiche autonome, una per ogni lingua considerata nel progetto (Farina 2016). Ogni squadra ha iniziato il proprio lavoro realizzando una bibliografia adeguata agli scopi del progetto, avendo come tematica i beni culturali nella propria lingua e le tipologie dei testi prima illustrate.

Inizialmente la squadra italiana ha effettuato una ricerca per costituire la bibliografia dei testi di maggiore rilievo che dovevano essere inseriti all'interno della banca dati, definiti testi fondatori o per la terminologia dell'arte (essendo l'arte italiana un punto di riferimento nella storia in generale) o perché sono classici dell'arte italiana (quali il Vasari o Michelangelo) aventi come argomento la città di Firenze.

Successivamente ogni squadra si è impegnata, oltre a cercare le traduzioni di tali testi, a individuarne altri nelle proprie lingue aventi come argomento l'arte a Firenze, ovvero libri di viaggiatori passati da Firenze (es. Stendhal) o manuali artistici di ampio valore culturale (ad esempio la *Geschichte der Renaissance* di Burckhardt).

Una volta individuati i testi di interesse per il progetto, prima di introdurre nel proprio corpus ogni testo[11], i responsabili di ciascuna squadra hanno compilato un foglio Excel con i metadati relativi al testo. Tali informazioni contribuiscono alla decisione del nome univoco da assegnare al file contenente il testo, al fine di facilitare il lavoro nella fase di consultazione delle concordanze; tale nome è "intellegibile", ovvero si compone di una serie di elementi che consentono

[10] In una successiva fase del lavoro si pensa di superare questo sistema di testi di riferimento e relative traduzioni per passare a creare corpus di riferimento del lessico artistico indipendenti in ognuna delle lingue. Ad esempio, per quanto riguarda la lingua spagnola è già attiva una collaborazione con la *Academia de Bellas Artes de San Fernando* per disporre di una banca dati dei testi fondatori del patrimonio artistico spagnolo che serva per lo studio del lessico dell'arte in spagnolo.

[11] Per poter consentire inizialmente ai componenti delle varie squadre di avere accesso ai vari file, si è deciso di utilizzare una soluzione di *storage* dei dati su *cloud*. Tutti i testi presenti all'interno della raccolta sono stati memorizzati come file Word (ovvero con estensione .docx).

di facilitare la corretta individuazione del testo oggetto di interesse, distinguendolo facilmente dagli altri presenti nella banca dati.

Contemporaneamente, dal punto di vista informatico, il lavoro è proseguito attraverso la realizzazione di un apposito script Python avente come scopo la riconciliazione di ogni testo con i relativi metadati e il successivo inserimento all'interno di un file XML; quest'ultimo ha subito quindi una procedura di tokenizzazione e di lemmatizzazione ad opera del lemmatizzatore TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/); i file "verticali" così ottenuti sono stati quindi inseriti in corpora all'interno di una installazione di NoSketchEngine su un server per uso interno. In tal modo sono state rese possibili ricerche di concordanze utilizzando filtri in base a varie caratteristiche, quali l'autore, il periodo storico, la tipologia testuale, ecc….

6. Analisi del testo

6.1 Aspetti generali

Avendo previsto un numero di utenti così ampio, è necessario che la possibilità di accesso per la consultazione dell'informazione sia quanto più ampia possibile; pertanto, per l'interrogazione alla banca dati si è scelto di utilizzare un analizzatore molto usato tra i membri del gruppo di ricerca, SketchEngine.

Le concordanze e la presenza di un opportuno linguaggio di *query* sono considerati caratteristiche basilari per la ricerca lessicografica e anche linguistica. Inoltre l'analizzatore prescelto offre la funzionalità di *word sketch*, rendendo questo utile per la creazione di lemmari di termini specifici[12] delle diverse arti o dei diversi componenti delle realtà artistico-turistica oggetto di studio.

Infine i *word sketch* possono essere utilizzati per mettere in relazione parole che si riferiscono a uno stesso oggetto artistico, per delimitare campi di conoscenza o di riferimento come monumenti e possono anche servire per differenziare oggetti da persone e disambiguare i nomi propri che fanno riferimento a queste due categorie. Per esempio in spagnolo e altre lingue l'uso delle preposizioni *a* e *en* possono indicare rispettivamente se il complemento diretto è di persona o il complemento circostanziale è di luogo.

[12] Interessanti riflessioni sulla estrazione di altre terminologie sono contenute in Bonin *et al*. (2010).

6.2 Aspetti tecnici

I software per la gestione dei corpora possono essere divisi in due categorie: quelli utilizzabili su computer desktop e quelli disponibili come *web service*. Per soddisfare le finalità del progetto di disporre di un ambiente utilizzabile ovunque, si è deciso di ricorrere a quei software utilizzabili come *web service* (e pertanto accessibili via internet).

Una ricerca dello stato dell'arte ha individuato l'esistenza di tre principali software all'interno della famiglia sopra descritta: CWB (http://cwb.sourceforge.net/), INL BlackLab (https://github.com/INL/BlackLab) e SketchEngine (https://www.sketchengine.co.uk/).

Occorre notare che, mentre i primi due software sono disponibili come *open source*, SketchEngine è disponibile solo dietro pagamento; tuttavia viene distribuita anche una versione gratuita ridotta, che va sotto il nome di NoSketchEngine, e che può essere liberamente utilizzata come *open source*, in cui sono assenti alcune delle funzioni della versione commerciale più interessanti per lo studio lessicografico, quali la funzionalità di *word sketch* ed il thesaurus.

Tutti e tre i software citati dispongono di una serie di caratteristiche di base di fondamentale importanza nell'ambito della ricerca delle concordanze, tra le quali: possibilità di effettuare ricerche semplici, ricerche mediante l'uso di espressioni regolari o di *query* CQL (Context Query Language, un linguaggio di Query nato originariamente per CWB ed implementato in seguito anche negli altri software citati), oltre alla possibilità di utilizzare software di lemmatizzazione quali TreeTagger, Freeling, RfTagger, ecc… Inoltre tutti e tre i software dispongono di una apposita API (Application Programming Interface) che consente di effettuare operazioni quali *query* o inserimento dati attraverso software terzi appositamente realizzati per le necessità della presente ricerca.

Mentre CWB e SketchEngine dispongono ormai di un'ottima documentazione e di una solida *community* di riferimento, che possono essere utilizzate con profitto nel corso delle fasi di implementazione, lo stesso non si può dire per BlackLab, che si dimostra ancora debole in tali aspetti.

In definitiva SketchEngine è stato preferito per la presenza di funzionalità interessanti come i *word sketch* e di una interfaccia utente elegante, con la possibilità di creare contemporaneamente un'installazione NoSketchEngine di supporto.

7. Primi risultati

Alla fine di dicembre 2017 la banca dati contava oltre sei milioni di parole, suddivise per lingua come indicato in Tabella 1.

Tabella 1. Numero di parole per lingua, a dicembre 2017

Lingua	Parole
Francese	2.885.000
Inglese	480.000
Italiano	870.000
Russo	440.000
Spagnolo	1.020.000
Tedesco	615.000

Inoltre, erano presenti 1229 testi, di cui 29 costituiti da libri interi e i restanti da frammenti "autoconclusivi".

In particolare, per quanto riguarda francese e spagnolo (le due lingue che hanno superato il milione di parole) è possibile evidenziare la distribuzione per tipologia e tematica di testo mostrata in tabella 2 (viene indicato il numero di testi presenti).

Tabella 2. Numero di testi per tipologia e tematica di testo, a fine dicembre 2017

Classificazione	Francese	Spagnolo
Divulgativo-Blog	1	0
Divulgativo-Guida	4	0
Divulgativo-Ricettario	0	2
Divulgativo-Rivista	1	0
Dizionario-Monolingue	6	0
Dizionario-Plurilingue	0	0
Letterario-Biografico	115	77
Letterario-Fiction	9	0
Letterario-Saggistica	14	3
Tecnico-Architettura	1	4
Tecnico-Arte	66	106
Tecnico-Edilizia	0	0
Tecnico-Enogastronomia	1	0
Tecnico-Storia	0	10

Infine, dall'interrogazione dei testi presenti nella banca dati mediante *query* CQL sono stati ottenuti i primi lemmari relativi al linguaggio dell'arte, di uso interno.

8. Problematiche e sviluppi futuri

Esistono alcune questioni che, data la breve esperienza di questo percorso, non sono state ancora verificate; in particolare, come descritto nel seguito, alcune scelte non sono state ancora consolidate come valide, e si auspica che l'utilizzo dei diversi corpora farà sì che possa essere confermata la loro validità, grazie alle problematiche che probabilmente dovranno essere affrontate nel percorso di raccolta dei testi in tutte le lingue.

Poiché la banca dati è stata progettata sulla base dell'esperienza di altri corpora generali[13] e specifici[14] è ipotizzabile che l'utilizzo dei corpora permetta di verificarne la sua validità; questo non è ancora avvenuto, poiché fino a questo momento sono state effettuate solo ricerche su corpora di piccole dimensioni, al fine di mettere in evidenza le concordanze.

La catalogazione dei testi per quanto riguarda la tipologia si è rivelata una fonte di possibili problemi. La distribuzione dei testi su epoche diverse si rivela un fattore origine di incertezze; ad esempio *Le vite* del Vasari è stato considerato un testo di tipo *Tecnico – Arte* e non *Letterario – Biografico* proprio tenendo in considerazione lo stato dell'arte della letteratura all'epoca del Vasari, ovvero per evitare una errata visione acronica dei testi abbiamo stabilito che la catalogazione dipenda dal carattere originale del testo nel momento di pubblicazione.

Inoltre, si è deciso di inserire all'interno della banca dati i testi come appaiono nella fonte oggetto di digitalizzazione, rimanendo quindi coerenti alla edizione che si sta copiando e alla grafia con cui erano stati scritti all'epoca della pubblicazione del libro. Questa scelta non è sempre stata accettata bene dalle varie squadre linguistiche, poiché può rendere difficili le ricerche su tutti i testi; nonostante questo si è preferito utilizzare tale modalità di lavoro poiché garantisce fedeltà al testo originale ed evita il rischio di possibili modernizzazioni con criteri non uniformi. Gli esperti delle diverse lingue potranno fornire delle tabel-

[13] Sono stati presi come riferimento i corpora dell'Accademia della Crusca e della Real Academia Española.

[14] Sono state prese in considerazione come modello del presente lavoro altre banche dati aventi argomenti simili, quali: la banca dati della Fondazione Memofonte e il Vocabolario Toscano dell'arte del disegno Opera di Filippo Baldinucci pubblicato dalla Scuola Normale Superiore di Pisa.

le di riferimento con i cambiamenti ortografici verificatisi nel tempo, e che potranno ad esempio essere utilizzate per facilitare l'esecuzione di ricerche su tali testi all'interno di SketchEngine.

Un aspetto che invece si rivela valido nei nostri corpora è che, sebbene non siano paralleli, tuttavia i testi sono stati inseriti in modo da poter effettuare tale operazione in un secondo tempo; questo fatto è di importante valore culturale e potrà essere molto utile per gli studi traduttologici. Inoltre, poiché si tratta di traduzioni di testi che hanno un carattere fortemente culturale, come i testi di Vasari o Stendhal (parlando della propria cultura o di una cultura diversa) potranno essere effettuati studi di carattere traduttologici e culturali di grande interesse.

Per il futuro è prevista la possibilità di effettuare la taggatura dei testi ad un livello molto più avanzato rispetto a quanto sia stato fatto finora, attraverso l'utilizzo di standard universalmente riconosciuti. In particolare, poiché la nostra banca dati è fortemente relazionata con i nomi propri di persona e di luogo, è ipotizzabile che uno degli strumenti adeguati sia ISNI; tale standard, applicando una etichetta ad ognuno dei nostri nomi permetterebbe di radunare i nomi propri in una lingua a tutte le varianti di un'altra lingua, ma anche di radunare in un solo riferimento tutte le varianti dei nomi propri, presenti nel corpus sia nella stessa lingua che tra lingue diverse. Ad esempio nel caso di Niccolò Machiavelli si potrà avere: Niccolò di Bernardo dei Machiavelli, Nicolás Maquiavelo, Nicolas Machiavel, Никко́ло Макиаве́лли, 尼科洛·马基雅弗利, ecc…

9. Considerazioni finali

Attualmente l'accesso alla banca dati è riservato esclusivamente ai membri del gruppo di ricerca, tuttavia se ne prevede nei prossimi mesi l'apertura a una comunità scientifica ristretta sia per effettuarne la validazione che per continuare il processo di sviluppo. È previsto ed auspicabile che vi sia una crescita e ristrutturazione della banca dati, tenuto conto anche del fatto che si prevede di continuare lo sviluppo per quanto riguarda l'area geografica italiana; invece, per quanto riguarda la tematica dell'arte, essa sarà mantenuta rigidamente senza oltrepassare un campo tanto vasto come quello dei beni culturali.

References

Accademia della Crusca. http://www.accademiadellacrusca.it/it/link-utili/banche-dati-dellitaliano-scritto-parlato (accessed November 4, 2017).

Bonin, F., Dell'Orletta, F., Montemagni, S. & Venturi, G. 2012. Lessico settoriale e lessico comune nell'estrazione di terminologia specialistica da corpora di dominio. In S. Ferreri (ed.), *Lessico e Lessicologia. Atti del XLIV congresso internazionale di studi della società di linguistica italiana: Viterbo, 27-29 settembre 2010*. Roma: Bulzoni, 207-220.

Carpi, E. 2017. El lenguaje para fines artísticos: traducciones de tondo al español. In A. Curado (ed.), *LSP in Multi-disciplinary contexts of Teaching and Research. Papers from the 16th International AELFE Conference*, 79-84.

Cresti, E. & Panunzi, A. 2013. *Introduzione ai corpora dell'italiano*. Bologna: Il mulino.

Farina, A. 2016. Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique. *Publif@rum* 25.

Fondazione Memofonte. http://memofonte.accademiadellacrusca.org/ (accessed November 4, 2017).

Glossario dei termini artistici. https://www.finestresullarte.info/glossario.php (accessed November 4, 2017).

ISNI. http://www.isni.org/ (accessed November 4, 2017).

McEnery, T. & Hardie, A. 2012. *Corpus Linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Ramos, M. 2009. Hacia un nuevo recurso léxico:¿ fusión entre corpus y diccionario?. In P. Cantos Gómez & A. Sánchez Pérez (eds), *A survey of corpus-based research*. 1191-1207.

Real Academia Española. http://corpus.rae.es/creanet.html (accessed November 4, 2017).

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing.

Tran, M. & Maurel, D. 2006. Prolexbase. Un dictionnaire relationnel multilingue de noms propres. *TAL* 47:115-139.

Vocabolario Toscano dell'arte del disegno. Opera di Filippo Baldinucci pubblicato dalla Scuola Normale Superiore di Pisa. http://baldinucci.sns.it/ (accessed November 4, 2017).